

## Supplementary data

### CpG domains downstream of TSSs promote high levels of gene expression

Simone Krinner<sup>1</sup>, Asli P Heitzer<sup>1</sup>, Sarah D Diermeier<sup>2</sup>, Ingrid Obermeier<sup>1</sup>, Gernot Längst<sup>2,\*</sup> and Ralf Wagner<sup>1,\*</sup>

<sup>1</sup> Department of Molecular Microbiology & Gene Therapy, Institute of Medical Microbiology and Hygiene,  
University of Regensburg, Germany

<sup>2</sup> Department of Biochemistry III, Institute for Biochemistry, Genetics and Microbiology,  
University of Regensburg, Germany

\* To whom correspondence should be addressed:

Dr. Ralf Wagner, Univ. Prof.  
Department of Molecular Microbiology and Gene Therapy  
Institute of Medical Microbiology and Hygiene  
University of Regensburg  
93053 Regensburg  
Tel: +49 941 944 6452  
Fax: +49 941 944 6455  
Email: [ralf.wagner@klinik.uni-regensburg.de](mailto:ralf.wagner@klinik.uni-regensburg.de)

\* Correspondence may also be addressed to:

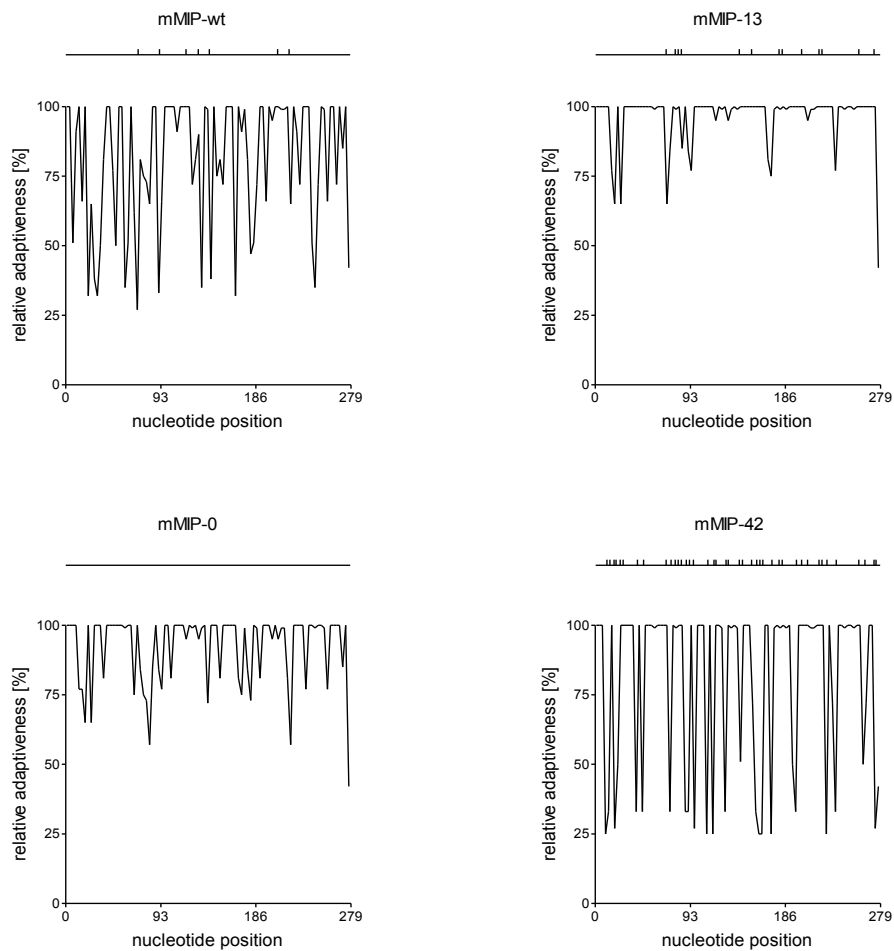
Dr. Gernot Längst, Univ. Prof.  
Department of Biochemistry III  
Institute for Biochemistry, Genetics and Microbiology  
University of Regensburg  
93053 Regensburg  
Tel: +49 941 943 2849  
Email: [gernot.laengst@ur.de](mailto:gernot.laengst@ur.de)

File contains supplementary figures 1-6 with figure legends and references.

**A**

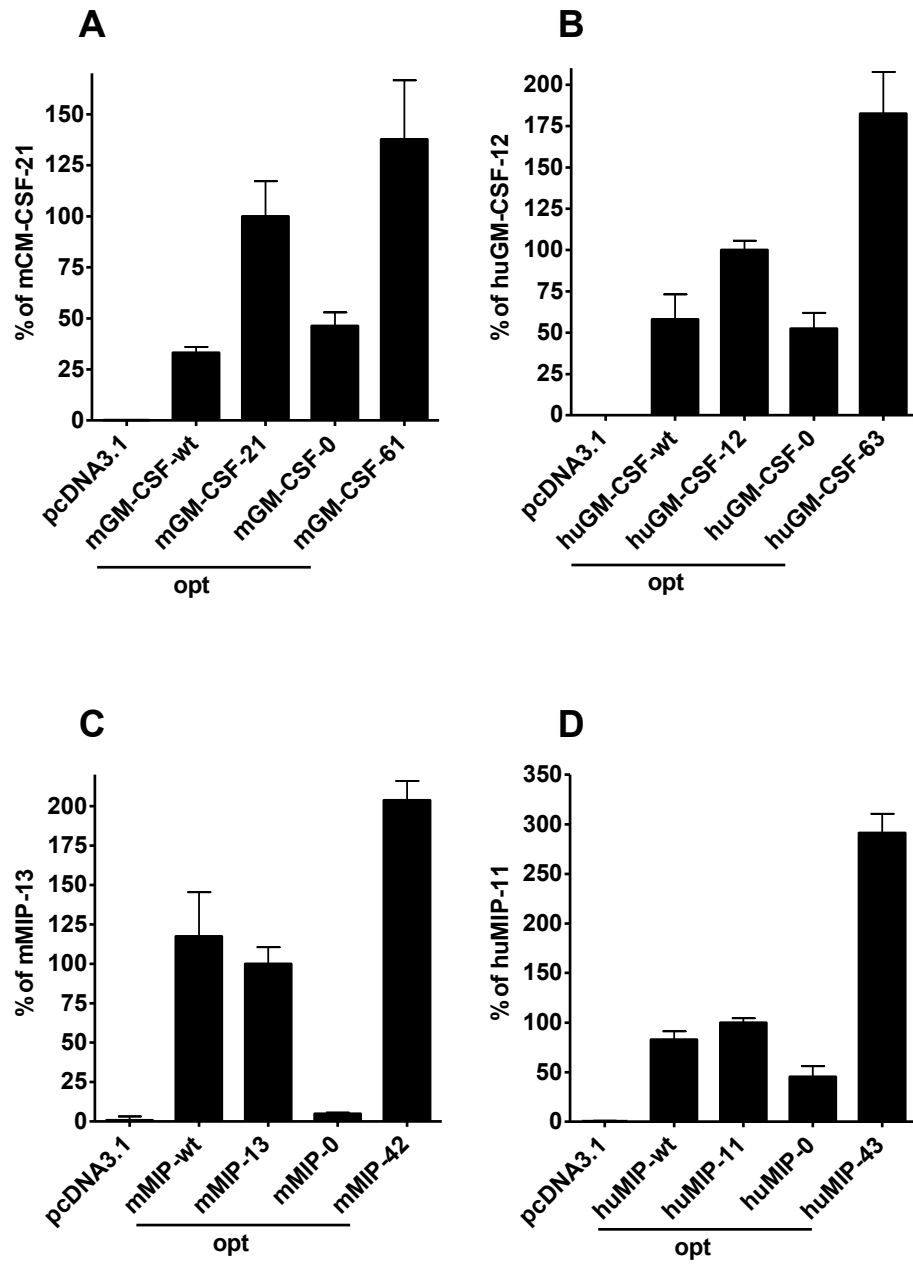
MURINE MIP-1A WT (7CPG)	ATGAAGGTCTCCTACACTGCCCCCTGGCTGTTCTTCTCTGTACCATGACACT	50
MURINE MIP-1A 0CPG	ATGAAGGTGAGCACAACAGCTCTGGCTGTCTCTCTGTACCATGACCT	50
MURINE MIP-1A 13CPG	ATGAAGGTGAGCACCACAGCTCTGGCTGTCTCTCTGTGACCATGACCT	50
MURINE_MIP-1A_42CPG	ATGAAGGTCTCAGCACCAGCGCTCGCCGTCTGCTGTGCACGATGACGCT	50
MURINE MIP-1A WT (7CPG)	CTGCAACCAAGTCTTCTCAGCGCCATATGGAAGCTGACACCCGACTGCT	100
MURINE MIP-1A 0CPG	CTGCAACCAAGTCTTCTCTGCCCTTATGGAAGCAGATACCCCTACAGCT	100
MURINE MIP-1A 13CPG	CTGCAACCAAGTCTTCTAGCGCTCCTTACGGCGCCGATACCCCTACAGCT	100
MURINE_MIP-1A_42CPG	CTGCAACCAAGTCTTCTAGCGCCCGTACGGCGCCGACACGCCGACCGCT	100
MURINE MIP-1A WT (7CPG)	GCTGCTTCTCTACAGCCGGAAGATTCCACGCCAATTCATGTTGACTAT	150
MURINE MIP-1A 0CPG	GCTGTTTTCAGTACAGCAGGAAGATCCCAGCCAGTTTCATTGTGACTAC	150
MURINE MIP-1A 13CPG	GCTGCTTCTAGTACAGCAGGAAGATCCCAGCCAGTTTCATCTGTGACTAC	150
MURINE_MIP-1A_42CPG	GCTGCTTCTCTGTTCTCGCGGAAGATCCCAGCCAGTTTCATCTGTGACTAC	150
MURINE MIP-1A WT (7CPG)	TTTGAACACGAGCAGCTTTGCTCCAGCCAGGTGTCATTTTCTGACTAA	200
MURINE MIP-1A 0CPG	TTTGAGACCAGCAGCCTCTGTTCTCAGCCTGGGGTCTCTTTCTGACCAA	200
MURINE MIP-1A 13CPG	TTTGAGACCAGCAGCCTCTGTTCTCAGCCGGCGTGATCTTCTGACCAA	200
MURINE_MIP-1A_42CPG	TTTGAACAGTCGTCGCTGTGCTCGCAGCCGGCGTGATCTTCTGACCAA	200
MURINE MIP-1A WT (7CPG)	GAGAAACCGCCAGATCTGCTGACTCCTAAAGAACCTGGGTCCAAGAAT	250
MURINE MIP-1A 0CPG	GAGGAACAGCCAGATCTGTGCAGACAGCAAGAGACATGGGTCCAGGAGT	250
MURINE MIP-1A 13CPG	GCGGAACAGACAGATCTGCGCCGACAGCAAGAGACATGGGTCCAGGAGT	250
MURINE_MIP-1A_42CPG	GCGGAACCGCCAGATCTGCGCCGACTCGAAGAAACGTGGGTCCAGGAGT	250
MURINE MIP-1A WT (7CPG)	ACATCACTGACCTGAACTGAATGCTTAG	279
MURINE MIP-1A 0CPG	ACATCAGACCTGAGCTGAATGCTTAG	279
MURINE MIP-1A 13CPG	ACATCAGACCTGAGCTGAACGCTTAG	279
MURINE_MIP-1A_42CPG	ACATCAGACCTCGAATCTGAACGCGTAG	279

**B**

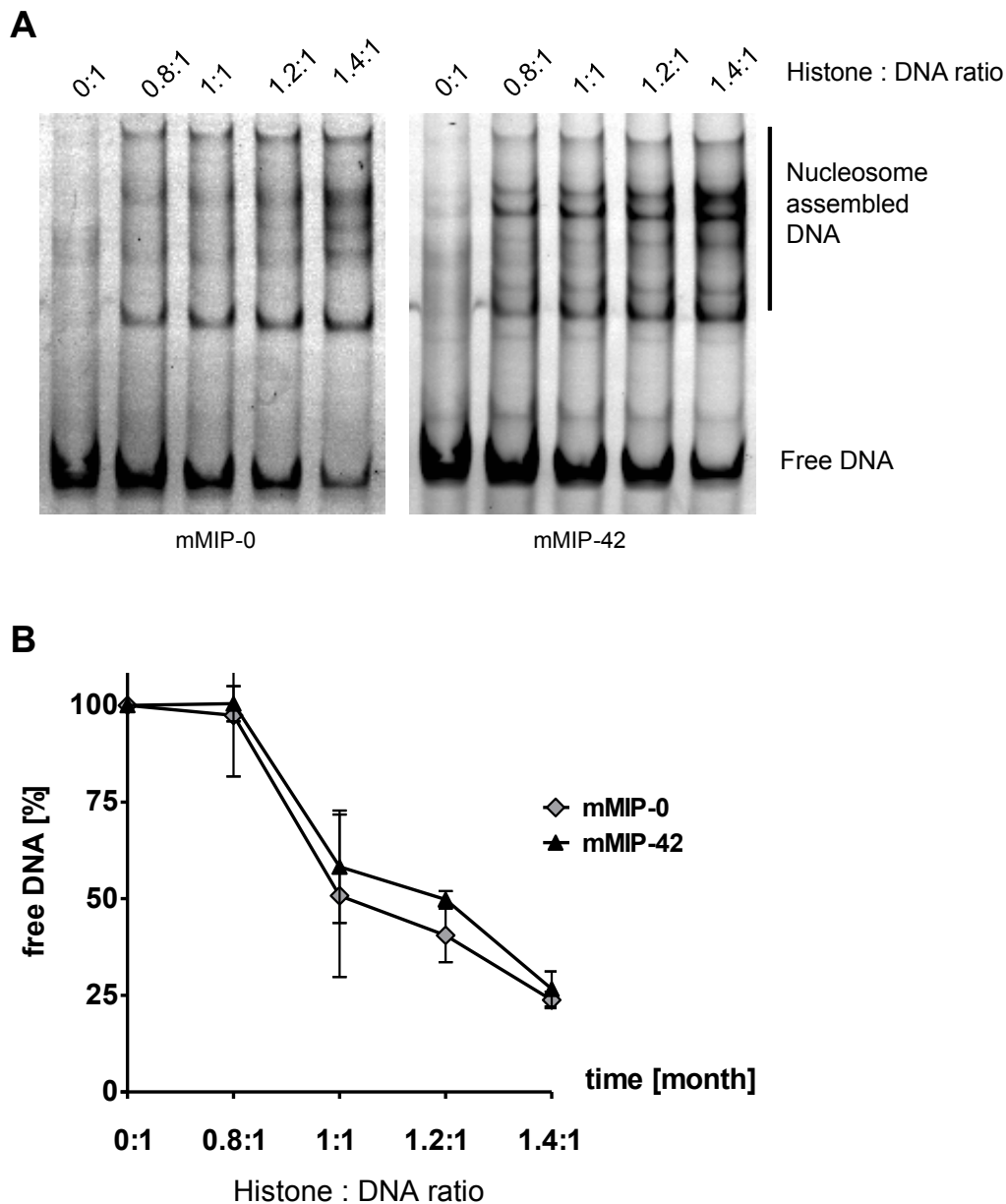


**Supplementary Figure 1 - *mmip-1a* variants and their relative adaptiveness distribution.**

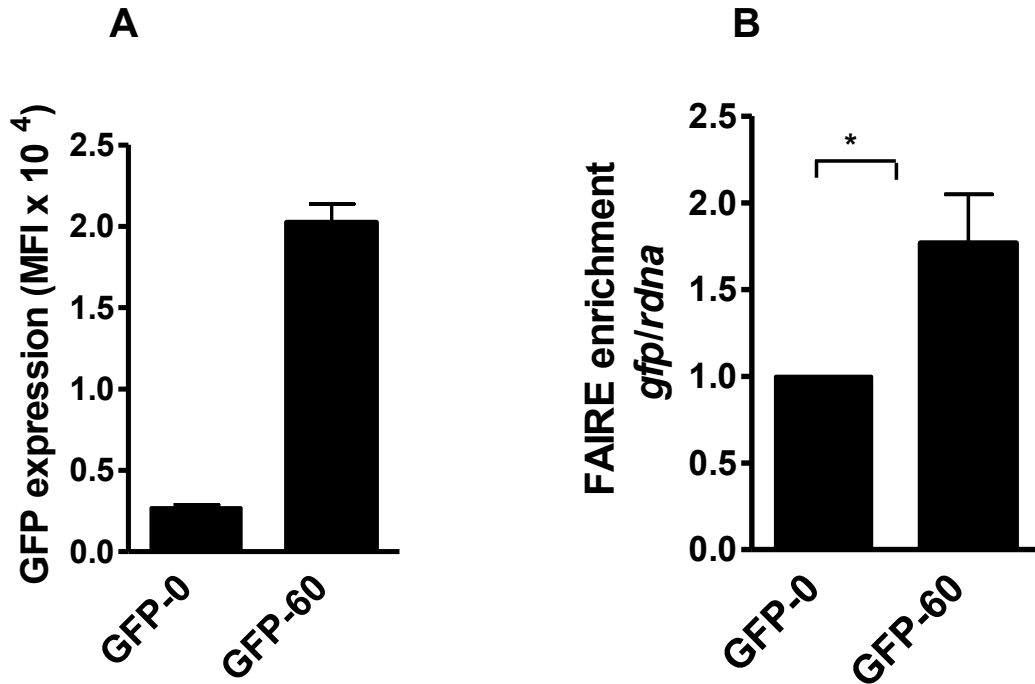
(A) Multiple nucleic acid sequence alignment of the open reading frames (ORFs) of *mmip-wt*, *mmip-13*, *mmip-0*, and *mmip-42*. Sequences were aligned using the DNAMAN software. The overall homology is indicated with colors (white: 100% homology; orange: 50% homology; blue: < 33% homology). (B) Schematic depiction of the *mmip-1α* variants and their relative adaptiveness distribution. Number and distribution of CpG dinucleotides within the ORF of *mmip-1α* (279bp) of the sense strand is shown. CpG dinucleotides are depicted as vertical lines. The relative adaptiveness reflects the frequency of individual codons, where the most frequently used triplet encoding a given amino acid is set to 100% and less frequently used codons are scaled down accordingly. The geometric mean of the respective relative adaptiveness results in the CAI of a given gene (1).



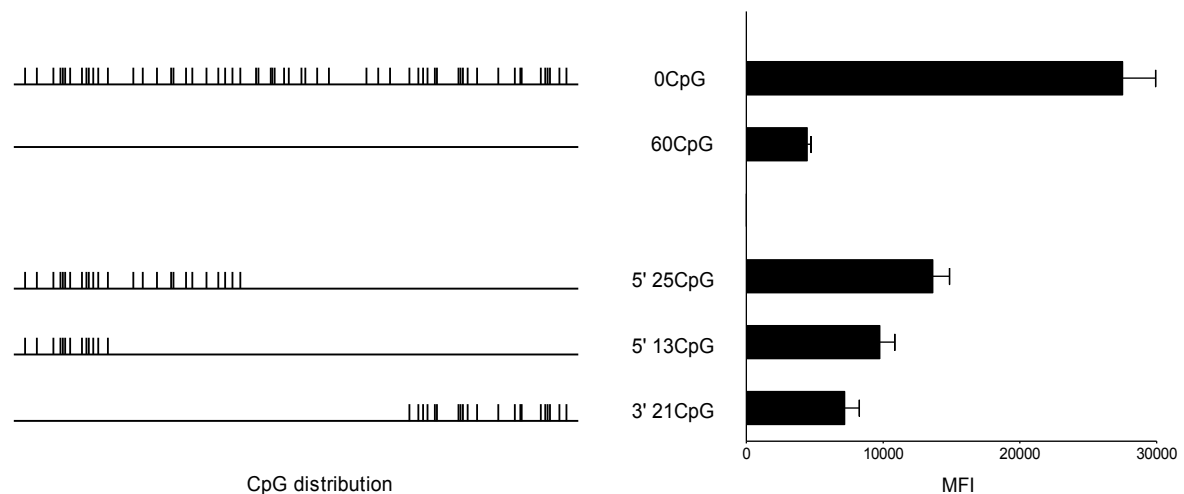
**Supplementary Figure 2 - Gene expression analysis of different cytokines in transiently transfected H1299 cells.** Quantification of mMIP-1 $\alpha$  levels by ELISA in the cell culture supernatants of H1299 cells transiently transfected with the gene variants of (A) *mgm-csf*, (B) *hugm-csf*, (C) *mmip-1 $\alpha$*  and (D) *humip-1 $\alpha$* . All gene variants are controlled by the CMV promoter. Mean and standard deviations of triplicates are shown.



**Supplementary Figure 3 - Influence of intragenic CpG dinucleotides on nucleosome affinity *in vitro*.** (A) Fluorescently labeled PCR fragments of *mmip-0* (DY550) and *mmip-42* (DY647) were reconstituted to mononucleosomes by salt dialysis, followed by PAGE and detection by fluorescence imaging. (B) Saturation levels of histone octamers with *mmip-0* and *mmip-42* fragments were quantified from free remaining DNA of decreasing histone:DNA ratios, measured by the Multi Gauge software. The intensity of optical density (IOD) of free DNA in mononucleosome reconstitutions was compared to the histone-lacking sample (100%). As judged from the comparison of decreasing IODs between gene variants, similar binding affinities for *mmip-0* and *mmip-42* were obtained. The mean and standard deviations of two assemblies are shown.

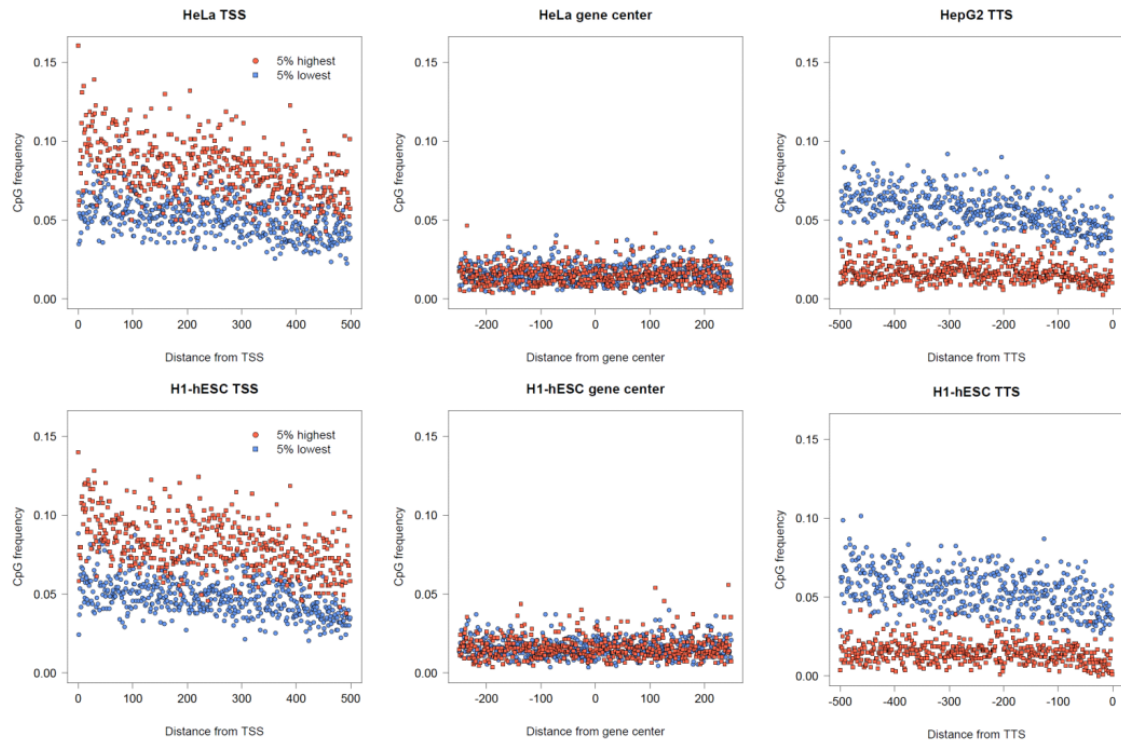


**Supplementary Figure 4 – Analysis of *hgfp* CpG variants stably transfected in CHO Flp-In cells regarding gene expression and chromatin density *in vivo*** (A) The gene variants *hgfp-0* and *hgfp-60* solely differ in their intragenic CpG content (no CpGs and 60 CpGs, respectively, adopted from preceding studies (2)). GFP expression of CHO Flp-In cells stably expressing the respective gene variant driven by the CMV promoter was assayed by flow cytometry. The mean fluorescent intensity (MFI) of hGFP positive cells is shown. The mean and standard deviations of three measurements is shown. (B) Enrichment for nucleosome-depleted chromatin by FAIRE extraction was performed, and DNA from the aqueous phase was quantified by real-time PCR using primer pairs specific for a region of the *gfp* ORF (+32 to +152 relative to the start codon). The values are presented as the ratio of DNA recovered from cross-linked cells divided by the amounts of the same DNA in the corresponding non-cross-linked samples. All results were normalized to *rdna* and referred to *gfp-0*, which was set to the value 1. The data reflect the degrees of nucleosome depletion in the respective genomic regions. The mean and standard deviations of two FAIRE preparations with a duplicate measurement each are shown. Significance was calculated using ANOVA/Tukey's Multiple Comparison Test (\*  $p < 0.05$ ).

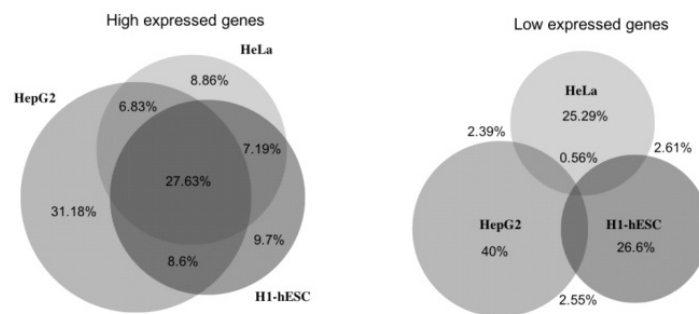


**Supplementary Figure 5 - Expression analysis of *gfp* chimera in stably transfected CHO Flp-In cells as quantified by flow cytometry.** On the basis of *gfp-0* and *gfp-60* (2), *gfp* chimera were generated by fusion PCR (left), followed by stable transfection into CHO Flp-In cells (right). Upon stable transfection of *gfp* chimera, GFP expression was analyzed by flow cytometry. The mean and standard deviations of triplicates are shown.

**A**



**B**



**Supplementary Figure 6 - Genome-wide correlation of CpG frequency and expression levels.**

(A) CpG frequency of the 5% highest and 5 % lowest expressed genes in HeLa and H1-hESC cell lines. Frequencies are exemplarily displayed within the first 500 bp, starting from the transcription start site (TSS, left), +/- 250 bp around the gene center (middle) and the last 500 bp of all genes, ending with the transcription termination site (TTS, right). Every symbol indicates the CpG frequency at the corresponding position. At the TSS, the CpG frequency of high expressed genes is up to 2-fold higher compared to low expressed genes. The occurrence of CpG dinucleotides decreases towards the gene centre and stays at a low level for high expressed genes, while low expression correlates to increased CpG values around the TTS. (B) Venn diagrams display the overlap of high and low expressed genes of the three different cell lines H1-hESC, HepG2 and HeLa. While the overlap of high expressed genes is about one third (27.63%) there is almost no compliance of the low expressed genes (0.56%). In absolute numbers, 1029 genes for each high and low datasets were assayed for H1-hESC, 1451 for HepG2 and 977 for HeLa.

### Supplementary references

1. Sharp,P.M. and Li,W.H. (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
2. Bauer,A.P., Leikam,D., Krinner,S., Notka,F., Ludwig,C., Langst,G. and Wagner,R. (2010) The impact of intragenic CpG content on gene expression. *Nucleic Acids Res*, **38**, 3891–3908.